Information Systems 37 (2012) 294–305

Contents lists available at SciVerse ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

An ontology-based retrieval system using semantic indexing

Soner Kara, Özgür Alan, Orkunt Sabuncu, Samet Akpınar, Nihan K. Cicekli*, Ferda N. Alpaslan

Department of Computer Engineering, METU, Ankara, Turkey

ARTICLE INFO

Available online 22 September 2011

Keywords:
Semantic web
Keyword-based
Query
Semantic indexing
Ontology

ABSTRACT

In this paper, we present an ontology-based information extraction and retrieval system and its application in the soccer domain. In general, we deal with three issues in semantic search, namely, usability, scalability and retrieval performance. We propose a keyword-based semantic retrieval approach. The performance of the system is improved considerably using domain-specific information extraction, inferencing and rules. Scalability is achieved by adapting a semantic indexing approach and representing the whole world as small independent models. The system is implemented using the state-of-the-art technologies in Semantic Web and its performance is evaluated against traditional systems as well as the query expansion methods. Furthermore, a detailed evaluation is provided to observe the performance gain due to domain-specific information extraction and inferencing. Finally, we show how we use semantic indexing to solve simple structural ambiguities.

© 2011 Elsevier Ltd All rights reserved.

1. Introduction

The huge increase in the amount and complexity of reachable information in the World Wide Web caused an excessive demand for tools and techniques that can handle data semantically. The current practice in information retrieval mostly relies on keyword-based search over full-text data, which is modeled as a bag-of-words. However, such a model misses the actual semantic information of the text. In order to deal with this issue, ontologies are proposed [1] for knowledge representation, which are nowadays the backbone of semantic web applications. Both the information extraction and retrieval processes can benefit from such metadata, which gives semantics to plain text.

Once the semantic knowledge is represented via ontologies, the next step is querying the semantic data, also known as semantic search. There are several query languages designed for semantic querying. Currently, SPARQL[1] is the state-of-the-art query language for the Semantic Web. Unfortunately, these formal query languages are not easy to be used by the end-users. Formulating a query using such languages requires the knowledge of the domain ontology as well as the syntax of the language. Therefore, Semantic Web community works on simplifying the process of query formulation for the end-user. The current studies on semantic query interfaces are carried out in four categories; keyword-based, form-based, view-based and natural language-based systems [2]. Among them, keyword-based query interfaces are the most user-friendly ones and people are used to use such interfaces easily, thanks to Google.

Combining the usability of keyword-based interfaces with the power of semantic technologies is one of the most challenging areas in semantic search. According to our vision of Semantic Web, all the efforts towards increasing the retrieval performance while preserving the user-friendliness will eventually come to the point of improving

* Corresponding author. Tel.: +90 312 2105582; fax: +90 312 2105544.
E-mail addresses: soner.kara85@gmail.com (S. Kara), alan@ceng.metu.edu.tr (Ö. Alan), orkunt@ceng.metu.edu.tr (O. Sabuncu), samet@ceng.metu.edu.tr (S. Akpınar), nihan@ceng.metu.edu.tr (N.K. Cicekli), alpaslan@ceng.metu.edu.tr (F.N. Alpaslan).

[1] http://www.w3.org/TR/rdf-sparql-query/.

0306-4379/$ - see front matter © 2011 Elsevier Ltd All rights reserved.
doi:10.1016/j.is.2011.09.004

semantic search with keyword-based interfaces. This is a challenging task as it requires complex queries to be answered with only a few keywords. Furthermore, it should allow the inferred knowledge to be retrieved easily and provide a ranking mechanism to reflect semantics and ontological importance.

In this paper, we present a complete ontology-based framework for the extraction and retrieval of semantic information in limited domains. The system consists of a crawler module, an automated information extraction module, an ontology population module, an inferencing module, and a keyword-based semantic query interface. Our main concern is creating a scalable and user-friendly information retrieval system with high retrieval performance. We applied the framework in the soccer domain and observed the improvements over classical keyword-based approaches. We show that our system achieves very high precision and recall values even for very complex queries in soccer domain. Furthermore, we evaluate and report the effects of different levels of indexing in terms of semantic processing (using only information extraction and using both information extraction and inferencing) on the query performance.

Scalability concerns are divided into two main topics; inferencing and querying. Scalability in terms of inferencing is assured by dividing the whole logical model into individual independent models since inferencing on a single large model is more complex than inferencing on independent smaller models. Similar studies are focused on representing the whole world in a single model. Therefore, they can not fit to large scales.

Scalability in terms of querying is assured by transforming the inferred knowledge into a single special inverted index structure. Using this inverted index structure scales our system up to web search engines, which means answering millions of queries in reasonable time and retrieving information from huge data sources. It also triggers the use of keyword based querying. In this way, the user-friendly way of querying is supported. Studies on ontology based information retrieval use logical queries on ontological models. Thus, their scale is restricted to small sized data sources compared to web scale and logical querying becomes a complex task for ordinary users.

Consequently, our main contribution is a framework which improves the performance of the keyword-based search using semantics without loosing the search scalability. In this aspect, this framework will be a strong base for ontology based search engines with its web scale crawler, information extraction, ontology population and inferencing modules.

The rest of the paper is organized as follows: A brief discussion about the related work is given in Section 2. In Section 3, we give the details of the components of the system, namely information extraction (IE), ontology population, inferencing and information retrieval (IR). In Section 4, we give the evaluation results. Section 5 gives a brief comparison with query expansion methods. Section 6 describes how the system can be extended to support phrasal expressions. In Section 7 we give a brief discussion and Section 8 concludes the paper with some remarks for future work.

2. Related work

The classical or traditional keyword-based information retrieval approaches are based on the vector space model proposed by Salton et al. [3]. In this model, documents and queries are simply represented as a vector of term weights, and the retrieval is done according to the cosine similarity between these vectors. Some of the important studies related to traditional search are [4–7]. This approach does not require any extraction or annotation phases. Therefore, it is easy to implement, however, the precision values are relatively low. The implementation of vector space models in real life applications is provided by the use of tools supporting the inverted index structures such as Lucene. In other words, Lucene like tools connect the real life applications to the theoretical background of vector space models.

The first step towards semantic retrieval is using WordNet synonym sets (synsets) for word semantics [8,9]. The main idea is expanding both indices and queries with the semantics of the words to achieve better recall and precision. If used together with an effective word sense disambiguation (WSD) algorithm, this approach is shown to improve the retrieval performance. On the other hand, a poor WSD will cause degradation in performance. Another drawback of this approach is the lack of complex semantics as it is limited to the relations defined in the WordNet.

Another step towards semantic retrieval is using information extraction. There are many studies on this field. Main dissimilarities between these studies arise from the structure of sources, details of the extracted information and computational/memory resources. NLP-based approaches are domain independent but use parse trees of sentences, pos taggers, chunk parsing, anaphora resolution, etc. in order to extract information. They need heavy computational processes [10–15]. There are some alternative information extraction methods such as pattern/rule-based information extractors against heavy computational costs. These methods are classified according to the creation forms of patterns and rules: automatic or manual. Automatic methods [11,16,17,13,18–21] are superior compared to the manual ones considering the effort spent on the domain. On the other hand, they suffer from low precision-recall rates.

The methods in [22–26] use hand-crafted rules to extract information. Hand-crafted rules are also used in semantic annotation [27–29]. Etzioni et al. [29] uses domain-independent rules to locate individuals of various classes in text. Wessman et al. [28] relies primarily on regular expressions. Cerno is a light-weight framework for semantic annotation of textual documents using domain-specific ontologies [27]. It combines keyword and structure-based annotation rules instead of linguistic patterns. We need a scalable information extractor for a rigid and structured domain. The extractor should also be appropriate for both Turkish and English content. Since the details of the extracted information is crucial for our purposes, we focus on high recall and precision values. Therefore, we have used a hand-crafted method whose details are given in [30].

Download English Version:

https://daneshyari.com/en/article/396555

Download Persian Version:

https://daneshyari.com/article/396555

Daneshyari.com